Classification & clustering

1. Consider the data set shown in Table:

Record	A	B	C	Class
1	0	0	0	+
2	0	0	1	-
3	0	1	1	-
4	0	1	1	-
5	0	0	1	+
6	1	0	1	+
7	1	0	1	-
8	1	0	1	-
9	1	1	1	+
10	1	0	1	+

a. Estimate the conditional probabilities for P(A|+), P(B|+), P(C|+), P(A|-), P(B|-), P(C|-). (For example, P(A=1|+)=3/5=0.6; P(A=0|+)=2/5=0.4).

b. Use the estimate of conditional probabilities given in the previous question to predict the class label for a test sample (A=0, B=1, C=0) using the naïve Bayes approach. **List the steps of prediction.**

2. Apply KMeans algorithm to cluster the following four objects (with (x, y) representing locations) into two clusters. Initial cluster centers are: Medicine A (1, 1) and Medicine B (2, 1). Use Euclidean distance. Explain each of the clustering steps and the clustering result after each iteration.

Object	Attribute 1 (x)	Attribute 2 (y)
Medicne A	1	1
Medicine B	2	1
Medicine C	4	3
Medicine D	5	4

3. Apply Naïve Bayes Classifier to classify the Iris data using R studio and compile the report. Use 10-fold cross validation, output and explain the classification result, including accuracy, precision, recall, F1 score, and confusion matrix.

4. Apply K-means to cluster the Iris data using R studio and compile the report.
   a. Set K from 2 to 6, report the internal index (SSE, BSS, and SC) for each K. Use SC to pick the optimal K's value. Visualize the clustering result using plot() function.
   b. Set K=3 and evaluate the result using the external index (i.e., the confusion matrix with the flower label).